Introduction to Explainable AI

Deep Learning Tutorial

2025-12-11

Padma, Ben, Franco, Luis. (Group 4)

Contents

1 xAI in Public Policy

Motivation: Why XAI?

As technology progresses, transparency is essential for ethical AI deployment.

The core of this idea lies in *human interpretability*:

- **"Interpretability is the degree to which a human can understand the cause of a decision" (Briran & Cotton, 2017)**

We need to interpret the *methods* and *predictions* behind AI models (Molnar, 2025), in order to trust and govern AI systems, especially when deployed in society, such as...

- COMPAS recidivism tool
- Medical triage algorithms
- Automated eligibility systems

Regulation is catching up:

- OECD guidelines and the EU AI Act demand clear explanations, bias checks and human oversight for high-risk systems
- Tradeoff: Performance vs Interpretability?

Case 1: COMPAS recidivism tool

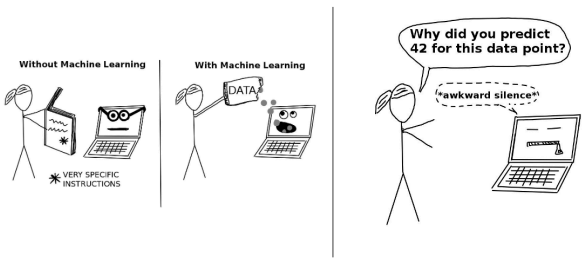
- Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS
- Tool used in U.S. courts to predict likelihood of reoffending
- **Controversy:** Alleged racial bias in predictions
  - A ProPublica investigation revealed that Black defendants were more likely to be incorrectly labeled as risky, despite not re-offending
- Model was proprietary and opaque
- Highlighted need for transparent and accountable AI-based decision making systems

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Figure 1: Example of Risk Prediction and Recidivism in Florida, source: ProPublica

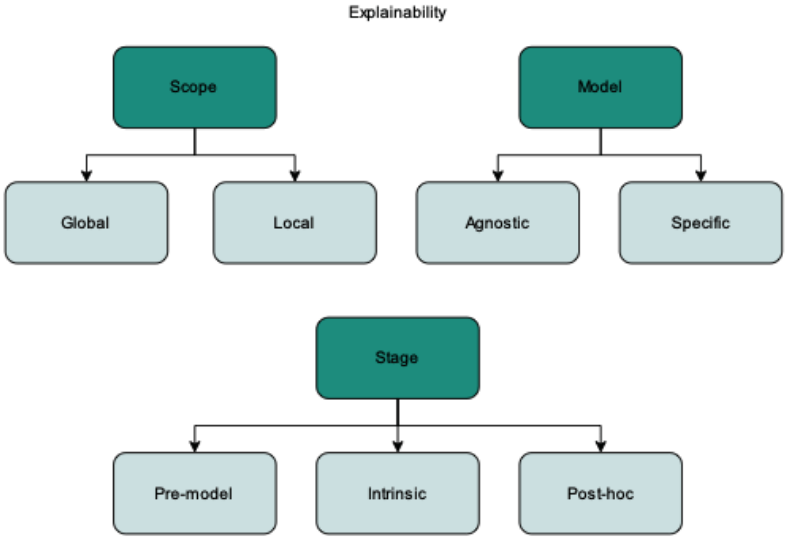
Case 2: Credit Scoring

- Credit scoring agencies use statistical models to evaluate creditworthiness
- **Controversy:** Lack of transparency in how scores are calculated
  - Consumers are often unaware which factors influence their scores
  - Individuals might be affected in their ability to obtain loans, housing, or employment
- Regulators stress the need for explainability to ensure fairness, prevent discrimination
- **Example:** EU's General Data Protection Regulation (GDPR) includes a *"right to explanation"* for individuals affected by automated decision-making
- ...yet enforcement and practical implementation remain challenging



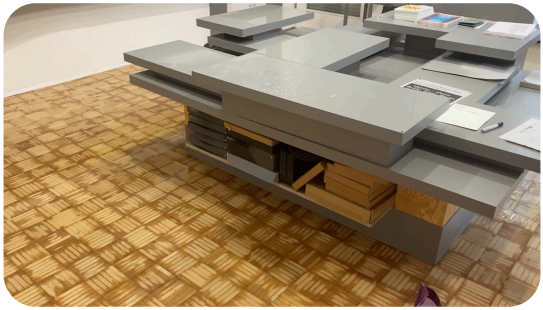
2 Methods

Taxonomy


Explainability
Scope
Model
Global
Local
Agnostic
Specific
Stage
Pre-model
Intrinsic
Post-hoc

Our Case

“I’ve always paid my loans back on time — what’s going on?”

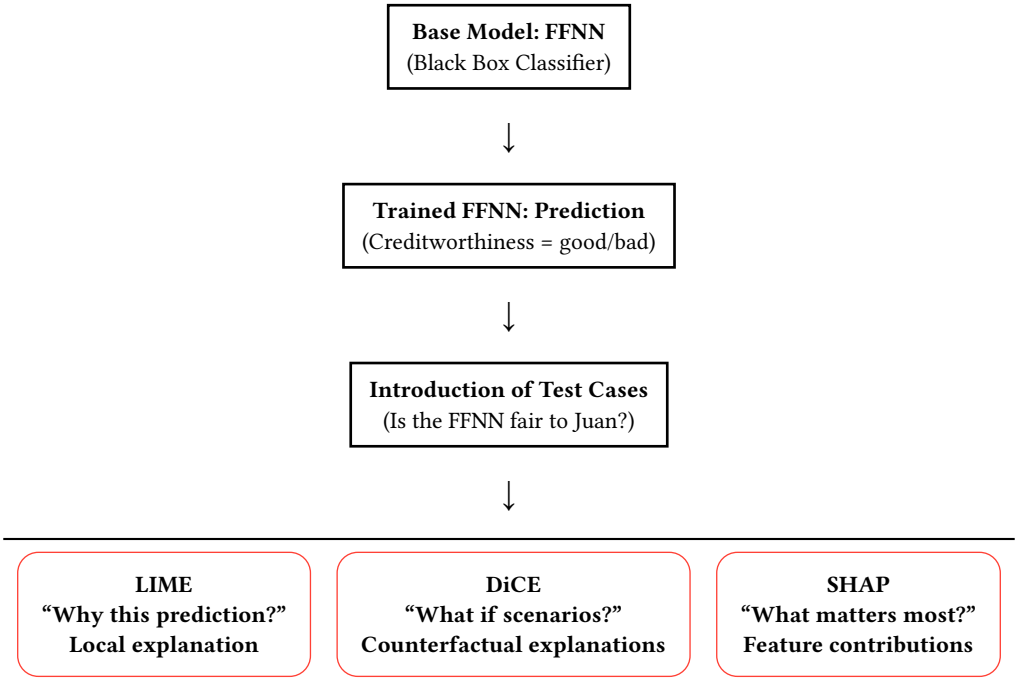


Meet Juan, a 35-year-old immigrant living in Germany.

He runs a small bookstore in Pankow, Berlin. Recently, flooding from an adjacent building damaged his shop, so he urgently applied for a loan to repair the property.

On paper, Juan looks like a strong applicant, but his loan gets denied. Can XAI methods tell us where the model failed him?

Toy Model: FFNN and Methods



LIME

"Why was Juan classified as high risk and therefore declined as a credit applicant?"

What is LIME?

- Local Interpretable Model-Agnostic Explanations
- Provides selective, local explanations for individual predictions

Why is it relevant?

- Shows why a single feature drove a specific decision
- Good to zoom in on an individual case and the model's behavior around that feature

DiCE

"What changes in Juan's feature profile would flip the decision?"

What is DiCE?

- Diverse Counterfactual Explanations
- *"What if"* scenario analysis into the features the model treats as most actionable

Why is it relevant?

- Counterfactuals map the applicant's profile and make choices more transparent
- Provide a way forward if we were to make model adjustments

SHAP

“Which features matter the most overall, across all combinations?”

What is SHAP?

- SHapley Additive exPlanations
- If features were players in a *game*, how much each would contribute to the overall payout, or prediction?

Why is it relevant?

- Provides global, game theory explanations of feature importance
- Detects feature interactions and nonlinearities relevant for deep learning

3 Takeaways

Takeaways

- At core: Human interpretability & oversight
- One should prioritize inherently interpretable models first
- If performance is critical, use of black box models should be accompanied by rigorous evaluation of explainability techniques
- Explainability methods can be useful but are limited; one needs to be cautious about their interpretations

4 Q&A

Acknowledgements

References

- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against Blacks. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- Molnar, C. (2025). Interpretable machine learning: A guide for making black box models explainable (3rd ed.). https://christophm.github.io/interpretable-ml-book/
- TransferLab. (n.d.). Explainable AI training materials. https://transferlab.ai/trainings/explainable-ai/

Template & Images

- typst template by diatypst
- Bookstore image: provided
- Blackbox image: Molnar, 2025

Annex I: LIME

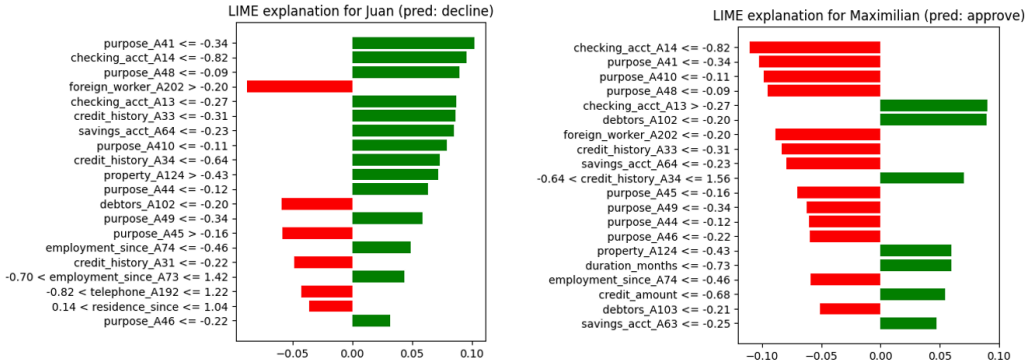


Figure 2: Local Explanations of FFNN Predictions: The Cases of Juan and Maximilian

Annex II: DiCE

```
JUAN'S LOAN APPLICATION

Juan's predicted class: decline

Juan's Application Data:
   duration_months  credit_amount  installment_rate  residence_since  age  existing_credits  maintenance_people  checking_acct_A12  checking_acct_A13  checking_acct_A14
0               14           3000                 2                4   35                 2                   1                  1                  0                  0


What Juan needs to change to get approved:
100%|██████████| 1/1 [00:00<00:00,  5.42it/s]

Found 2 scenario(s) with changes to foreign_worker, age, credit_amount, or duration:

Scenario 1:
  Columns that need to change:
            credit_amount  foreign_worker_A202
current            3000.0                  1.0
scenario_1        11895.7                  0.0

Scenario 2:
  Columns that need to change:
            duration_months  credit_amount
current                14.0         3000.0
scenario_2             39.6        16411.3
```

Figure 3: Counterfactual Analysis based on Scenarios 1-3 of Potential Feature Changes

Annex III: SHAP

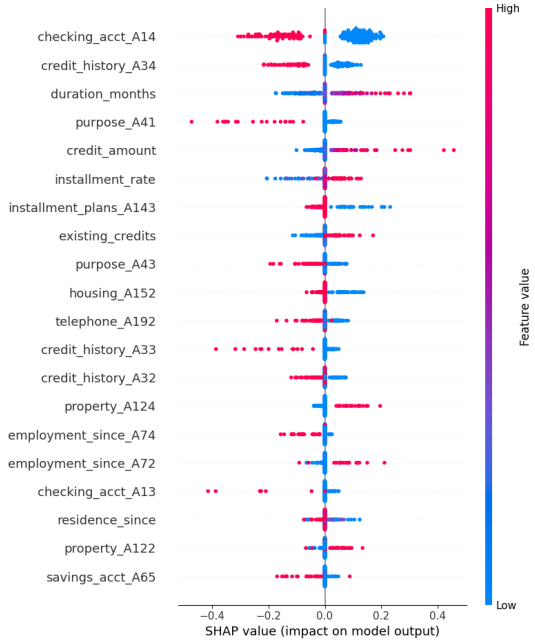


Figure 4: Global Feature Importance based on SHAP